JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
Volume 03 Issue 01
2022

Research Article

A Feature Fusion Based Hybrid Approach for Breast Cancer Classification

Fatima Iftikhar[1], Hafiz Muhammad Mueez Amin[2*] and Ghulam Abbas[3]

[1]Department of Computer Science, National College of Business Administration and Economics, Lahore, Pakistan
[2]Department of Computer Science, MNS-University of Agriculture, Multan, Pakistan
[3]Department of Computer Science, Virtual University of Pakistan, Lahore, Pakistan
[*]Corresponding Author: Hafiz Muhammad Mueez Amin. Email: 2014-uam-350@mnsuam.edu.pk

Received: December 11, 2021 **Accepted:** February 21, 2022 **Published:** March 15, 2022

Abstract: A detected type of cancer is breast cancer commonly in women. According to some estimate one in nine women is diagnosed with breast cancer. It is unfortunate that due to a lack of proper facilities, the diagnosis of breast cancer in patients is being delayed, which is leading to an increase in the possible death rate. Many different statistical methods and Machine Learning algorithms are often employed in the study to make breast cancer detection more accurate. Machine learning (ML) has allowed doctors to achieve remarkable results, and healthcare is using ML-based models to detect breast cancer in women. This allows analyzing the healthcare data and uses the traditional computer-aided detection (CAD) to assess breast cancer. Machine learning has become an accepted clinical practice and allows doctors to evaluate the ML model to detect breasts at an early stage. A major aim is to diagnose patients with breast cancer by analyzing the data of patients and classifying them into two categories, having diagnosis results as Benign "B" or Malignant "M". In this study different machine learning algorithms are used to classify cancer as either its malignant or benign. The Kaggle data set was used for applying these algorithms to get the best accuracy. MLP is more efficient and accurate algorithm to classify the breast tumor. And here also fitted the matthews_corrcoef for MLP is 0.89% and accuracy score for the random forest is 0.94%.

Keywords: Breast cancer; Machine learning; Breast tumor; Classification

1. Introduction

Cancer is the leading cause of death among women in the world. Breast cancer deaths are rising alarmingly around the world. Breast cancer is the most common cancer in women worldwide. According to the survey in 2018, the speed of breast cancer in 2.4 million women can be estimated from the fact that one out of every 4 women suffers from cancer [1], [2]. Breast cancer is lower rates in Asian Countries than in Western Countries. But over time, the rate of breast cancer has increased in Asian countries as well [2]. According to recent estimates, the ways cancer patients are being registered in Pakistan one in nine women suffer from this problem. Pakistan ranks first among Asian Countries where the rate of cancer in women is increasing with age [3]. It is unfortunate that due to a lack of proper facilities, the diagnosis of breast cancer in patients is being delayed, which is leading to an increase in the possible death rate [4]. If breast cancer is firstly approached by scientific method then, the survival rate of Pakistani women can be significantly improved [5]. There is no repository or database related to any disease also as breast cancer in Pakistan, but only the hospital is referred for data, which includes the number of new cancer patients and their annual death rate [6]. The use of breast cancer screening techniques in modern times has shown that the rate of cancer in women varies in Asian countries and Western countries [1].

The breast cancer in the women of Western countries has been increased in women aged 50 and over. In developing countries, on the other hand, younger women have been found to have 47% more breast cancer than older women. Breast cancer as usually found in women in their 40s and 60s in the Asian population. Breast cancer patients in Asian countries, especially India, Korea, and Japan, range in age from 40 to 49 or 50 to 55 years [7]. Breast cancer is a type of disease that starts from out-of-control cells in breast tissues. It develops anywhere in the body by the crowd of our cells. It makes it difficult for the body to work. Mostly it starts from a tumor or the growth of the tumor. Sometimes a lump called a tumor, but not all the lump is cancer [8]. Which piece of the lump is taken out to find out its cancer is named a biopsy. Lumps that do not cause cancer are called benign. Lumps that spread the cancer dangerously in the body are called malignant. A complex group of diseases that have many possible symptoms like, Genetics, Environmental Factors, Lifestyle habits, Carcinogens sometimes there are not obvious causes. Medications often help to reduce or kill cell growth. Every drug is different working or used together to treat the cancer patient. The abnormal growth of cancerous breast tumors is not spread outside of the breast. All the treatments are not permanent, but it also defines in many sessions such as 8 to 12 onward. Most breast cancers start in different parts of the breast.one of them is Ductal cancer which starts from the nipples that carry milk is named ductal cancer, the second cancer that occurs in the glands that makes milk in the breast is called lobular cancer. A small number of breast cancers that start in the tissues of are called Sarcomas & Lymphomas. A less common type of breast cancer is Phyllodes tumor and Angiosarcoma [9]. Not at all the causes of breast cancer are lumps in a breast. Breast cancers are also found on screening mammograms. It detects at the early stages of cancer before its symptoms are developed.

Many women come in for regular screening or checkups and examination of breasts. About 74% of women go for routine screening in addition to the DM. The radiologists take into account the final assessment category that is based on DM. Israeli women who are aged 50-74 are given free screening services for breast cancer. When DM has to be performed, including future and early screening processes will require. This indication appeared for around 15.5% of women. In this way, the results will be clear so the different situations are distinguished while the suspicions are excluded. Women between the ages of 45 to 50 are more likely to get the disease than other women. If they get regular screening it will become easy for them to detect the disease and get treatments at the right time. The screening policy in many countries is free and women can avail these services if they plan to take care of their health in the future [10]. As women age they take the self-examination for breast cancer, this is what they discussed in the questionnaires too. If you have a family member or relative with breast cancer it is more important to get regular screening to avoid any bad happenings. Women who are into smoking and drinking are also at risk of this disease.

1.1 Risk factors

The risk factor of breast cancers is necessarily meaning you will develop breast cancer. Many women develop breast cancer with unknown risk factors. Other than, if you are women, this in itself is a very high risk of getting cancer. If you have a biopsy before, then your risk of breast cancer increases. A chance of increasing your age, you are more likely to get breast cancer. Mainly we see the diagnosed breast cancer particularly at a young age if your sister, mother, and daughter were diagnosed. Meanwhile, many women have no family history of this disease. In fatty women, we see obesity is a major cause of cancer. The Menopause Cycle before the age of twelve or at an older age also increased in breast cancer. These women have also shown signs of cancer that have never been pregnant after the age of 30 except those women who had one or more pregnancies at this age. Hormonal therapy of estrogen and progesterone to treat the symptoms of menopause also cause the major risk factor of breast cancer so it is an important way to aware woman to stop these medications. Other factors of breast cancer are drinking more alcohol, tobacco, and other types of drugs [11]. Women who have been diagnosed with cancer of the colon, ovary, or Endometrium increases the risk of breast cancer. When your symptoms appear, talk to a qualified doctor or a close friend or even a close person to your family. At the right time, with the help of a doctor's advice, the risk of breast cancer can be reduced. The fat tissues in the bodies are the main source of stop producing hormones in ovaries. Having fat tissues increased the level of higher estrogen. After the study of breast and relationships between diets are suggested that it also reduce its risk. A very low-fat diet maintains the risk factor. Estimating and measuring the weight of a healthy body can be done with BMS. Eat the foods full with omega-3 and fatty acid avoids processed meat and Tran's food [12].

1.2 Survivor of Cancer

A cancer survivor is a type of person who is still living. Every survivor of cancer has its individuals, concerns, and challenges. A good first step is being able to recognize your fear, shyness and talk about them. They find support from friends or family members, your healthcare team, individual counseling, and the support from the place where you receive your treatment. A survivor may experience a great mixture of fear, guilt, sorrow, concern, and strong feeling. The two main types of survivors are firstly the people who appreciate their life after they diagnosis it and accept their disease. Secondly, others become anxious uncertain about the challenges and pain they face. They may feel much stress with their visits to healthcare and their relationship built them. The people miss their source of support. It is a bitter reality that a survivor also faces worries and challenges over time such as emotional challenges, any late effects of treatment, fear of recurrence, sexual health, fertility concerns, and financial issues [13]. Over the last 55 years, we see the ratio of increasing cancer patients. According to a 1971 survey, 3 million people have cancer. Today that number has grown from 3 million to 15.5 million. About 67% of people get survived each year. 17% of all cancer survivors are diagnosed 20 years ago. 47% of survivors are at the age of 70 or older [14]. It is the main responsibility of people to try to make their lives better, to be kind to them, cooperative with them, not to hurt them and to accept them with their weakness [15]. A process that is discovering useful information from a big dataset, functions, and data mining techniques helps to find any kind of disease, statistics, database, fuzzy sets, neural network, and warehouse help in the diagnosis of different cancer diseases [16] For example, lung cancer [17], leukemia [18] and prostate cancer. The detection of the traditional methodology of cancer is based on "the gold standard". There are three types of tests performed in data mining like pathology test, clinical examination, radiological imaging [19]. The presence of cancer in this procedure is regression process, the techniques and algorithms are based on a model design in new machine learning. In the proposed model the unseen data provides a good exact result on testing stages [20]. The main three stages of machine learning that will be applied to available datasets are preprocessing of data, selection of features or its extraction, and classification [21]. A major process that actually helps to predict the symptoms of cancer is called feature extraction. It elaborates cancer set by analyzing the data of patients and classifying them into benign and malignant tumors [22]. Using machine learning and data mining methods, we predict many types of breast cancer at an early stage. Regression and classification are the definite techniques [23], which have higher chances of predicting breast cancer in patients. By using the datasets, we trained our model and apply many algorithms to get more accurate kinds of breast cancer.

1.3 Why Machine Learning?

The machine learning algorithm can help doctors to detect breast cancer. The method uses the dataset of mammograms to detect cancer at its first stage. This is a definite method and has higher chances of early detection of the disease. The doctors view the electronic health records of the patient, and this helps them to make breast cancer detection more accurate. Machine learning has more accuracy as compared to radiology [24]. Digital mammography (DM) is the major method used to detect breast cancer. The screening for breast cancer involves the diagnostic and other health care data of the patient. Machine learning has been introduced in clinical practice, and the radiologist is evaluating mammograms using the techniques of machine learning. If an abnormal finding is detected during a diagnostic, then additional mammographic tests are carried out. The imaging modalities indicate the chances of risk of breast cancer in women. If a lesion or something suspicious is detected, then the machine learning methods are further used to analyze the risk and cause of cancer [25]. If something suspicious is found, then a biopsy is recommended. Analyzing the images can be challenging because there can be a subtle difference between lesions and the fibrous glandular tissue that indicates cancer. Different lesion types are present in a small proportion and might not be able to appear in the screening. The average probability of the positive test result allows measuring the chances of breast cancer. The radiologist has reported that the 86.9% sensitivity and the 88.9% specificity will indicate the risk of Breast cancer. The models can be identified based on clinical features that have allowed the physicians to estimate and identify the probability of women that can develop breast cancer. Machine learning (ML) has allowed doctors to achieve remarkable results, and healthcare is using ML-based models to detect breast cancer in women. This allows analyzing the healthcare data and uses the traditional computer-aided detection (CAD) to assess breast cancer. Machine learning has become an accepted clinical practice and allows doctors to evaluate the ML-DL model to detect breasts at an early stage [24], [25]. Combined machine and deep learning model that is based on the dataset. These are linked to mammograms and patient health records that improve the models in many ways. Many radiologists are available for screening purposes and it is an important part to detect cancer and start with the screening for breast cancer.

1.4 Improved Breast Cancer detection using Machine Learning Algorithms?

Machine learning has gained a lot of importance medically. Due to machine learning the results come fast and analyzing the situation will also become better. Radiologists can come up with better reports and treat patients will a lot more effort. Benign ones are not difficult to handle and the doctor may give some medications for it while the malignant ones can be dangerous and threatening [26]. If you look, machine learning is pretty much the same as data mining [40], [46].

1.5 Detection of Breast Cancer using Machine Learning Algorithms

Breast cancer is a disease that is increasing the death rate every year. It is one of the most common types of cancers found among women. Classification and data mining methods have been effective in detection previously, but it does not work in the field anymore. The false results could be a problem here so the medical industry has moved further with machine learning. In the medical field, a lot of diagnosis and analysis is used and some of them may not be appropriate either. Different sets of machine learning algorithms are taken into account for obtaining data. The algorithms include Support Vector Machine (SVM), Decision Tree (C4.5), Naive Bayes (NB), and k-Nearest Neighbors (KNN) [27]. The major objective of these algorithms is to find out the correctness of classifying the data using machine learning. It has to be checked how efficient, effective, precise, and sensitive it is for obtaining accurate results. When the experiment and data are taken, machine learning proves to be 97% successful in extracting results. There is no doubt that deep learning has improved the detection of this deadly disease. All the algorithms will bring in good results and the radiologists can get their reports to facilitate patients [28]. Detection of breast cancer can be highly challenging as the images on the screening mammography has to be observed diligently. The image classification is difficult as the tumor occupies only a small area of the screen. The image will be augmented and ROI annotations have been available in a wide variety for the mammogram databases. It will help establish detection and classification methods that include the region-based convolution neural network and many variants. Even large mammography databases can lack a lot of ROI annotations. It is laborious and costly to assemble all these annotations. There are very few public mammography databases that are fully annotated to secure the best of results. It is not easy to prove whether these annotations can bring favorable results as many challenges have to be faced by radiologists [29]. A deep learning machine requires a large range of training datasets for it to become even more effective. The classification algorithms will be seen to check the cancer status on each image. However, for the accuracy to occur it may not be very favorable. Pre-training is an important type of Machine learning to take place effectively. The machine learning Database has many images in terms while screening for the mammography all these elements will be taken into the picture [30]. A machine-learning algorithm allows doctors to predict breast malignancy, and the results come out in 12 months. The algorithm helps to identify breast cancer and has helped to detect 34 of 71 (48%) women. These women's initial radiologist was negative, but with machine learning, their cancer was detected in the year. The machine learning algorithm shows the white blood cells and also explains the functioning of the thyroid that indicates the risk of breast cancer.

A major aim is to diagnose patients with breast cancer by analyzing the data of patients and classifying them into two categories, having diagnosis results as: (1) Benign "B" (2) Malignant “M” based upon his/her tumor features i.e. its radius, area, smoothness, texture, and perimeter. Machine learning classification techniques can be used on this type of problem statement. Malignant: it’s too much cancer. If it occurs in one part of our body, it spreads quickly to another part. And once the doctor removes it, it comes back easily. Benign: it is not very cancerous. It does not spread quickly to other parts of the body, as it does to any other part of our body. And once the doctor removes it, it does not come back easily.

The paper structure is as follows: literature review is defined in the next section, methodology is defined in section 3, experimental results are defined in section 4, conclusion is defined in section 5 and references are given at last.

2 Literature Review

In this section, we have pointed out the prediction and extrapolation of subtypes related to breast cancer which smears a vital part in the identification and diagnosis of dangerous disease of the breast. In current years, deep learning frameworks, techniques, and methods exposed and revealed a well-intentioned enactment in the brainy and smart extrapolation of subtypes as well as breast cancer predictions. Conversely, the utmost of the outmoded and conventional deep learning models practiced particular modality data. Which could just have excerpted insufficient features? Thus, it could not establish and inaugurate a constant association in physiognomies of patients and subtypes.

In a study, the researchers utilized the dataset that is indicated as TCGA-BRCA used in the form of a mockup set for the subtype. Also, it inspected and predicted the molecular reasons behind breast cancer. After this, they used a dataset of TCGA-BRCA in detailed experiments with the deep learning model. In this paper, they established a methodology. They assembled a Hybrid dep learning model. On the other hand, they projected multimodal data. They united the data of gene modality of patients with the data of modality yet comprised of images. Based upon that, they fabricated a multi-modal synthesis framework.

The researchers claimed that the performance of computer-aided verdict in-short CAD for cancer could be promoted by modern improvements in deep learning combined with radionics mining high-level sorts and structures from the images of remedies in medical. They said that the cancer deformity of a breast is the most recurrent cancer among women and computer-aided structures. They also discerned that, based on deep learning approaches and procedures from images breast disease of re-counted potential accomplishments. In this paper, the researcher aimed to deliver an all-inclusive impression of the modern research exertions upon deep learning various radionics in the study area of the breast.

In order to determine innovative structures for breast corpus taxonomy, they smeared a framework comprised of convolutional neurons with supreme two convolutional layers and a maximum of two layers of pooling, and one layer termed as fully associated. In their framework of CNN, they showed an upsurge from seventy-nine percent to eighty-six percent in testing paralleled to the outmoded frameworks of radionics. In preceding researches on the improvement of deep learning frameworks and approaches for cancer prophecies, the approaches were undeveloped and artless. Thus, with the pervasive awareness and consumption of DL, however, more they planned altered designs with cavernous convolutional fed with images of cancer to advance the diagnosis process [33].

To generate the deep learning of lymph nodes with images of metastasis, ten deep learning structures were automatically designated in the crucial regiment. An area beneath the curvature of 0.8 in the prime regiment and 0.9 in the validation regiment was obtained. The anticipated deep learning model established for auspicious biased capacity. An additional signature model was established to advance distinguish the number of nodes of metastatic. In conclusion, a signature model for pre-functional extrapolation in images of metastasis position and quantities has been recognized for the patients suffering from cancer. Subsequently, deep learning dependent upon signatures might hypothetically afford an intrusive explanation to benefit clinicians in patients [35].

During normal clinical practice at the time of mammogram interpretation, the deep learning model assessment of mammographic breast density was introduced to the radiologist. The proportion of mammograms evaluated as dense by all radiologists decreased from 47.0 percent before the implementation of the deep learning model to 41.0 percent after the implementation of the deep learning model. As a result, their deep learning model had a high rate of clinical acceptance among both academic and community radiologists and decreased the share of dense mammograms assessed. This is often an important move before possible widespread deployment to validate their deep learning model [36].

Cancer may be a daunting health issue associated with high death worldwide. With the rapid advancement of high-throughput sequencing technologies and therefore the introduction of varied machine learning techniques that have developed in recent years, improvement has been gradually made in cancer prediction supported organic phenomenon, providing awareness into effective and precise decision-making in care. Therefore, there's considerable current interest in developing machine learning models which will accurately differentiate cancer patients from healthy people. However, nobody approach outperforms all the others among the classification methods applied to cancer prediction thus far. They demonstrated a replacement approach during this paper, which applies deep learning to an ensemble approach that integrates several different models of machine learning. They supplied five separate classification models with insightful gene data selected by differential organic phenomenon analysis.

A deep learning technique was then wont to assemble the outputs of the five classifiers. The proposed multimodel ensemble approach supported deep learning is shown to be accurate and efficient for cancer prediction by taking full advantage of various classifiers [37], [38].

2. Materials and Methods

Breast cancer is predicted in this paper using a variety of machine learning techniques. To fulfill this purpose, this paper is divided into two parts. The first part is to pre-process the data and the second part is to train the model for predicting Breast cancer. The dataset used to predict breast cancer is taken from Kaggle [39] where the data is available to the public so they can use it for their research. This dataset has 569 instances and 6 attributes. These six attributes are:

Table 1. Attributes of the Dataset

Attributes	Description	Range
Diagnosis	Diagnosis of Breast cells(1 = Malignant, 0 = Benign)	0-1
Mean_radius	Between the distance from the center to the points on the parameter	0-1
Mean_texture	standard deviation of Gray-scale values	0-1
Mean_perimeter	mean size of the core tumor	0-1
Mean_area		0-1
Mean_smoothness	mean of local variation in radius length	0-1

Data pre-processing is an important step before classification. Data pre-processing includes data cleaning, data dimensionality reduction, data transformation, data normalization, and data processing. Our data cleaning technique includes filling within the missing values if present with the mean of the attributes. We represent cancer cell detection using the Pearson coefficient of correlation as a feature selection technique. There are various algorithms [41], [42] that predict breast cancer. We have selected the six best classifiers Support Vector Machine (SVM), Decision Tree (DT), Naive Bayes (NB), k Nearest Neighbors (KNN), Random Forest, Logistic Regression, and MLP. After reading the research papers produced the best accuracy with the lowest error rate. We will evaluate all these classifiers to check which will provide more accurate results for our proposed methodology with the lowest error rates for the prediction of Malignancy of Breast cancer cells. Our proposed methodology is illustrated in Figure 1.

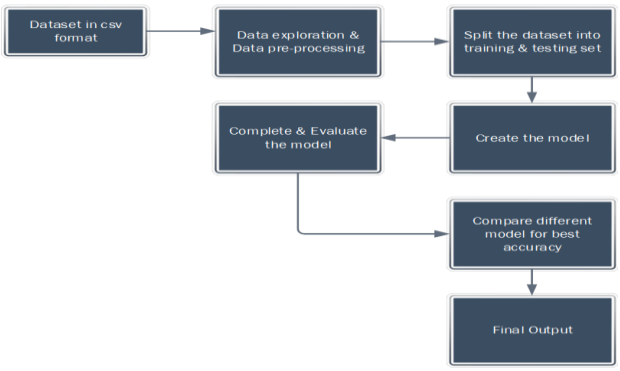

Figure 1. Our Proposed Methodology

This diagram describes all the steps from beginning to end, how to pre-process the data and train the model by using this data and train the model to classify Breast cancer.

4 Experimental Results

In this section analysis of data, data description and data pre-processing are discussed. We have used machine learning algorithms (ML) for the classification of the breast cancer either the cancer is malignant or benign. Here we have the basic steps of machine learning like exploring the dataset, Data pre-processing and cleaning then splitting and applying the model. First we have uploaded the data and read the data. Second we have applied preprocessing that is cleaning part shaking null values.

4.1 Import Essential libraries

We have used pandas with machine learning techniques for research. Firstly, we have imported our libraries then we used pandas to analyze the data, numpy for our mathematical calculation, seaborn and matplotlib for our data visualization.

4.2 Reading Breast Cancer dataset

We have made a dataset to understand and more readable form with the name mentioned as "df.head". It contains 569 rows and 6 columns. The attributes/columns names are Mean_radius, Mean_texture, Mean_perimeter, Mean_area, Mean_smoothness and diagnosis. There are total 569 out of 357 malignant patients and 212 benign patients.

4.3 Data Visualization

For data visualization pair plot is used for breast cancer data. In the pair plot, tumor is divided into two classes of malignant or benign. As you can see that the representation of malignant is 1 and benign is 0. Pair plot makes the data easier to understand.

4.4 Heat map of breast cancer dataset

The heat map makes it easy for us to understand the variety of features. In the right bar it represents that "higher the colors lower the valve". As in this heat map like if the color is light it diagnosis that the patient has malignant tumor or if the color is dark it represents that the patient has benign tumor.

4.5 Data Preprocessing

In this portion we have checked that there are any null values or not. As you see below in Figure 4 there are no null values in this dataset.

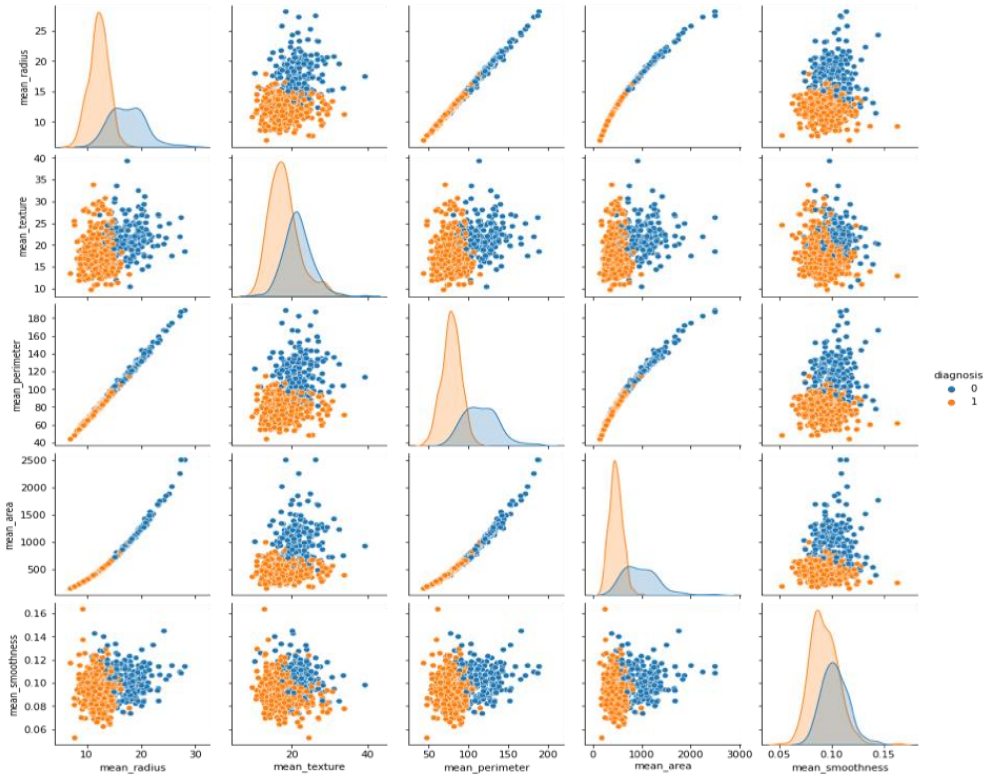

Figure 2. Pair plot of breast cancer dataset

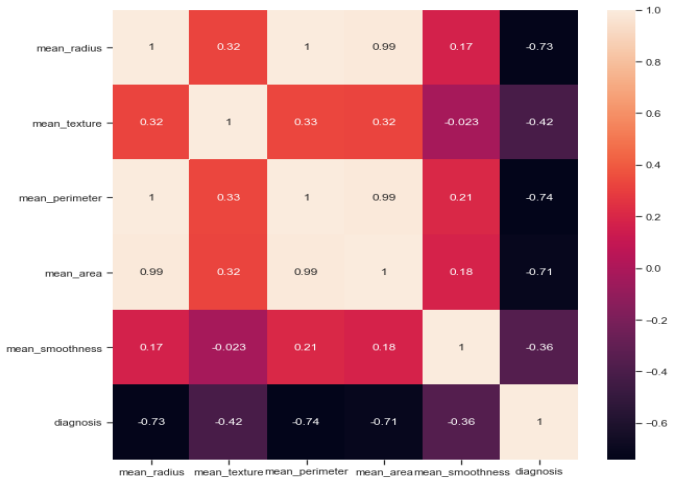

Figure 3. Heat map of breast cancer dataset

```
In [122]: dia_data.isnull().sum()
Out[122]: mean_radius          0
          mean_texture         0
          mean_perimeter       0
          mean_area            0
          mean_smoothness      0
          diagnosis            0
          dtype: int64
```

Figure 4. Null values in the dataset

```
In [121]: dia_data.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 569 entries, 0 to 568
Data columns (total 6 columns):
mean_radius          569 non-null float64
mean_texture         569 non-null float64
mean_perimeter       569 non-null float64
mean_area            569 non-null float64
mean_smoothness      569 non-null float64
diagnosis            569 non-null int64
dtypes: float64(5), int64(1)
memory usage: 26.7 KB
```

Figure 5. Data types of each attribute

For the information part, we have just checked that the data type and you can see that all the categories are in "float" expect the diagnosis which is in "int" part. The best model that provide the highest accuracy than other model and the parameters that can be used in measuring performance of the algorithms are discussed below.

4.6 Modeling Method

Six different types of classification models are used in this study that is the Decision Tree (DT), Logistic Regression (LR), Random Forest (RF), K-Nearest Neighbor (KNN), Support Vector Machine (SVM) and MLP. The main purpose of using these models is to give the best possible results for cancer prediction. Figure 6 describes the whole procedure of the model.

First, the CSV file of our dataset was pre-processed. And then it was classified to build its own model to predict either the tumor is malignant or benign. This data is then divided into two parts, one for training the data and the other for testing the data. After training the trained data set for better classification of the model, it was found that MLP is the best model for cancer prediction accuracy or its effectiveness. After getting better prediction results through MLP, we compared it with the other models, DT, LR, RF, KNN, SVM.

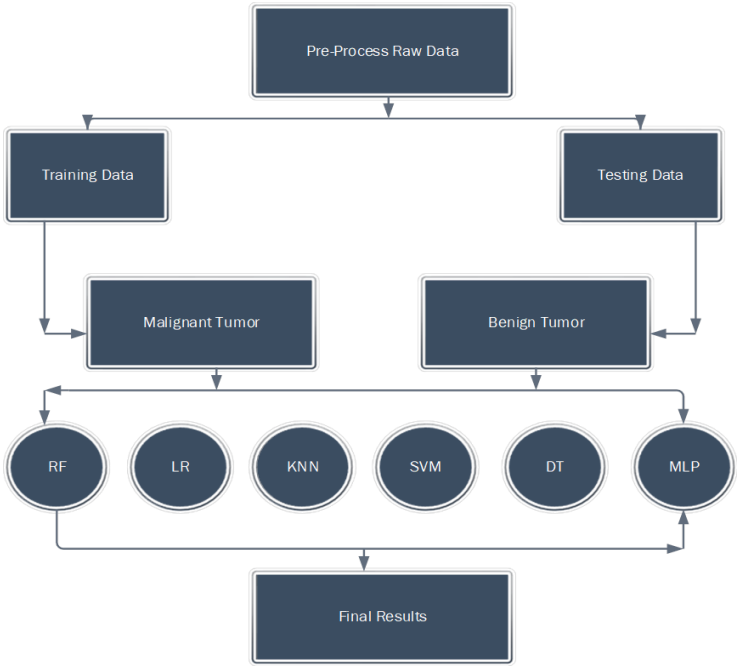

Figure 6. Process of Proposed Model

This section uses some parameters to measure the performance of six experimental classification models based on machine learning techniques. The confusion matrix used to measure these performance matrixes are True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN). In this study following parameters and formulas are used to measure performance.

Accuracy (Acc) Proportion of correct classification (true positives and negatives) from overall number of cases.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Precision is the number of correct positive results divided by the number of predicted positive results.

$$\text{Precision} = \frac{TP}{TP+FP}$$

Recall is the number of correct positive results divided by the number of actual positive results

$$\text{Recall} = \frac{TP}{TP+FN}$$

The **F-score** is the Harmonic mean of Precision and Recall.

$$F = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Percision} + \text{Recall}}$$

4.6.1 Classification report using Random Forest

Firstly, Random Forest model was used. Random forest is a kind of Ensemble classifier which is using decision tree algorithms [49]. And here also fitted the matthews_corrcoef for random forest is 0.80% and accuracy score for random forest is 0.90%.

Table 2. Classification report for Random Forest

	Precision	recall	f1-score	Support
0	0.89	0.88	0.88	48

1	0.91	0.92	0.92	66
micro avg	0.90	0.90	0.90	114
macro avg	0.90	0.90	0.90	114
weighted avg	0.90	0.90	0.90	114

4.6.2 Classification report using Logistic Regression

Secondly logistic regression model was used. Now from the name itself you might think that this is used for regression, but this is used for classification [47]. And here also fitted the matthews_corrcoef for logistic regression is 0.87% and accuracy score for random forest is 0.93%.

Table 3. Classification report for Logistic Regression

	Precision	**Recall**	**f1-score**	**Support**
0	0.94	0.92	0.93	48
1	0.94	0.95	0.95	66
micro avg	0.94	0.94	0.94	114
macro avg	0.94	0.94	0.94	114
-weighted avg	0.94	0.94	0.94	114

4.6.3 Classification report using KNN

Thirdly KNN model was used. It is very important algorithm for classification and regression that normally used in data mining. KNN is sufficient on working on small size of data that's why it is very easy to implement [50]. And here also fitted the matthews_corrcoef for KNN is 0.87% and accuracy score for random forest is 0.93%.

Table 4. Classification report for KNN

	precision	**Recall**	**f1-score**	**Support**
0	0.94	0.92	0.93	48
1	0.94	0.95	0.95	66
micro avg	0.94	0.94	0.94	114
macro avg	0.94	0.94	0.94	114
weighted avg	0.94	0.94	0.94	114

4.6.4 Classification report using SVM

Fourthly SVM model was used. SVM is a kind of supervised learning used for classification and regression analysis [42]. And here also fitted the matthews_corrcoef for SVM is 0.87% and accuracy score for random forest is 0.93%.

Table 5. Classification report for SVM

	precision	Recall	f1-score	support
0	0.94	0.92	0.93	48
1	0.94	0.95	0.95	66
micro avg	0.94	0.94	0.94	114
macro avg	0.94	0.94	0.94	114
weighted avg	0.94	0.94	0.94	114

4.6.5 Classification report using Decision Tree

Then decision tree model was used. Decision trees [43] are applicable in two different data mining techniques that are classification and prediction. It is used to visually define the rules which are simple to interpret and understand. Simply says it generates rules and we can visualize these rules by decision trees [44], [45]. And here also fitted the matthews_corrcoef for decision tree is 0.75% and accuracy score for random forest is 0.87%.

Table 6. Classification report for Decision Tree

	precision	Recall	f1-score	support
0	0.84	0.88	0.86	48
1	0.91	0.88	0.89	66
micro avg	0.88	0.88	0.88	114
macro avg	0.87	0.88	0.87	114
weighted avg	0.88	0.88	0.88	114

4.6.6 Classification report using MLP

Lastly MLP model was used. And here also fitted the matthews_corrcoef for MLP is 0.89% and accuracy score for random forest is 0.94%.

Table 7. Classification report for MLP

	precision	Recall	f1-score	support
0	0.94	0.94	0.94	48
1	0.95	0.95	0.95	66
micro avg	0.95	0.95	0.95	114
macro avg	0.95	0.95	0.95	114
weighted avg	0.95	0.95	0.95	114

4.7 Discussion

In the given below graph different algorithms random forest, logistic regression, KNN, SVM, decision Tree, and MLP are illustrated. In this graph three things about cancer prediction will be represented, one is accuracy second is F1_score and the third one is matthews_corrcoef. As their accuracy is mentioned on the below graph the best algorithm that can classify the tumor either it is malignant or benign is the MLP algorithm. MLP algorithm is one of the best algorithms accuracies or it can also provide the best MCC ratio rather than other algorithms.

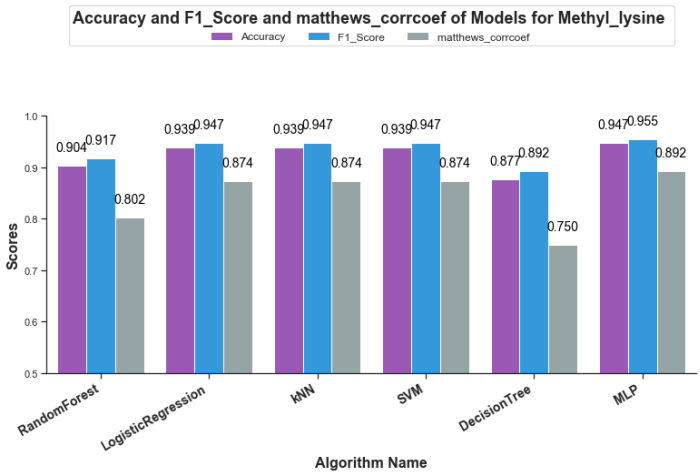

Figure 7. Best Accuracy Algorithm

In a random forest, it will provide 0.90% accuracy. Or in logistic regression, it will provide 0.93% accuracy. Or in KNN it will provide 0.93% accuracy. Or in SVM it will provide 0.93% accuracy. Or in a decision tree, it will provide 0.87% and lastly, MLP provides 0.94% accuracy. According to this percentage of accuracy, we suggest that MLP is best for classify the breast tumor.

5. Conclusions

There are many methods of data mining and machine learning to learn medical data. The most important issue in the area of data mining and machine learning is how to create a better classifier that is useful for medical science applications. The algorithms I have used in my research are six comparatively logistic regression, random forest, KNN, SVM, decision tree, and MLP.

The highest accuracy is given by the MLP algorithm which is 0.94%. In contrast, the lowest accuracy is given by the decision tree which is 0.83%. The main aim of using machine learning algorithms is to make it easier to detect tumors, while in the field of medical science it takes more time and more money to do the same things. Machine learning techniques work as a clinical assistant in any system for new doctors and physicians to diagnose breast cancer. MLP has proven to be the best of all the techniques to predict breast cancer. Further use of MLP can have amazing benefits in predicting cancer. At the end of this research, we suggest that machine learning techniques are able to predict any disease automatically more accurately.

5.1 Recommendations

Machine learning models are improving day by day and we can use them in different fields, it is also possible to use new and more advanced models to better compare the results. One of the machine learning models is that more layers and parameters the model has, more data results in better performance will be.

The sample size which was used for this study is small so it is recommended that the classification models can be impalement on large size data. These models can also be used on the whole tumor classification of the experimental base. It is also recommended that to use a different combination with the ML model.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre et al., "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," CA: A Cancer Journal for Clinicians, vol. 68, no. 6, pp. 394–424, 2018.
2. T. Vos, R. M. Barber, B. Bell, A. B. Villa, S. Biryukov et al., "Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990-2013: A systematic analysis for the Global Burden of Disease Study 2013," Lancet, vol. 386, no. 9995, pp. 743–800, 2015.
3. S. Sohail and S. N. Alam, "Breast cancer in Pakistan - Awareness and early detection," Journal of College of Physicians and Surgeons Pakistan, vol. 17, no. 12, pp. 711–712, 2007.
4. I. Karachi, P. F. Yasmeen and S. Zaheer, "Functional Time Series Models to Estimate Future Age-Specific Breast Cancer Incidence Rates for Women," Journal of Health Sciences., vol. 2, pp. 213–221, 2014.
5. E. Karadeli, G. Erbay, A. Parlakgumus and Z. Koc, "Utility of diffusion weighted magnetic resonance imaging with multiple b values in evaluation of pancreatic malignant and benign lesions and pancreatitis," Journal of College of Physicians and Surgeons Pakistan, vol. 28, no. 2, pp. 103–109, 2018.
6. M. Hanif, B. Sabeen, A. Maqbool, A. Ahmed, F. Nadeem et al., "Breast Cancer: Incidence (Thirteen Year Data Analysis) and One Year Clinicopathological Data of Patients in a Tertiary Care Cancer Hospital," International Journal of Biology and Biotechnology, vol. 12, no. 3, pp. 373–379, 2015.
7. E. H. Park, S. Y. Min, Z. Kim, C. S. Yoon, K. W. Jung et al., "Basic facts of breast cancer in Korea in 2014: The 10-year overall survival progress," Journal of Breast Cancer, vol. 20, no. 1, pp. 1–11, 2017.
8. C. D. C. Prevention, "What Is Breast Cancer? | CDC." [Online]. Available: https://www.cdc.gov/cancer/breast/basic_info/what-is-breast-cancer.htm. [Accessed: 21-Feb-2021].
9. Healthline, "Breast Cancer: Symptoms, Stages, Types and More." [Online]. Available: https://www.healthline.com/health/breast-cancer#types. [Accessed: 21-Feb-2021].
10. C. D. C. Prevention, "What Is Breast Cancer Screening? | CDC." [Online]. Available: https://www.cdc.gov/cancer/breast/basic_info/screening.htm. [Accessed: 21-Feb-2021].
11. A. Qaseem, J. S. Lin, R. A. Mustafa, C. A. Horwitch and T. J. Wilt, "Screening for breast cancer in average-risk women: A guidance statement from the American College of Physicians," Annals of Internal Medicine, vol. 170, no. 8, pp. 547–560, 2019.
12. C. D. C. Prevention, "What Are the Risk Factors for Breast Cancer? | CDC." [Online]. Available: https://www.cdc.gov/cancer/breast/basic_info/risk_factors.htm. [Accessed: 21-Feb-2021].
13. CancerNet "Breast Cancer: Survivorship | Cancer.Net." [Online]. Available: https://www.cancer.net/cancer-types/breast-cancer/survivorship. [Accessed: 21-Feb-2021].
14. A. B. Mariotto, R. Etzioni, M. Hurlbert, L. Penberthy and M. Mayer, "Estimation of the number of women living with metastatic breast cancer in the United States," Cancer Epidemiology, Biomarkers & Prevention, vol. 26, no. 6, pp. 809–815, Jun. 2017.
15. H. Sung, C. E. DeSantis, S. A. Fedewa, E. J. Kantelhardt and A. Jemal, "Breast cancer subtypes among Eastern-African–born black women and other black women in the United States," Cancer, vol. 125, no. 19, pp. 3401–3411, Oct. 2019.
16. M. K. Gupta and P. Chandra, "A comprehensive survey of data mining." International Journal of Information Technology, 12(4), 1243-1257, 2020.
17. D. Delen, "Analysis of cancer data: a data mining approach," Expert Systems, vol. 26, no. 1, pp. 100–112, Feb. 2009.
18. M. Shahbazl, S. Faruq, M. Shaheen and S. A. Masood, "Cancer Diagnosis Using Data Mining Technology," Life Science Journal, 9(1), 308-313, 2012.
19. A. R. Vaka, B. Soni and S. Reddy, "Breast cancer detection by leveraging Machine Learning," ICT Express, vol. 6, no. 4, pp. 320–324, Dec. 2020.
20. V. Sandri, I. L. Gonçalves, G. M. D. Neves and M. L. R. Paraboni, "Diagnostic significance of C-reactive protein and hematological parameters in acute toxoplasmosis," Journal of Parasitic Diseases, vol. 44, no. 4, pp. 785–793, Dec. 2020.
21. S. Eltalhi and H. Kutrani, "Breast cancer diagnosis and prediction using machine learning and data mining techniques: A review." IOSR Journal of Dental and Medical Sciences, 18(4), 85-94, 2019.
22. J. M. Wang, C. L. Qian, C. H. Che and H. T. He, "Study on process of network traffic classification using machine learning." In 2010 Fifth Annual ChinaGrid Conference (pp. 262-266). IEEE, July, 2010.
23. D. L. Olson and D. Delen, "Advanced data mining techniques." in Springer Science & Business Media, 2008.
24. X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks." in Proceedings of the thirteenth international conference on artificial intelligence and statistics (pp. 249-256), March 2010.
25. D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, 2015.
26. C. Meads, I. Ahmed and R. D. Riley, "A systematic review of breast cancer incidence risk prediction models with meta-analysis of their performance," Breast Cancer Research and Treatment, vol. 132, no. 2. pp. 365–377, Apr-2012.
27. Y. Wu, C. K. Abbey, X. Chen, J. Liu, D. C. Page et al., "Developing a utility decision framework to evaluate predictive models in breast cancer risk estimation," Journal of Medical Imaging, vol. 2, no. 4, p. 041005, Aug. 2015.
28. S. Sharma, A. Aggarwal and T. Choudhury, "Breast cancer detection using machine learning algorithms." In 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS) (pp. 114-118), IEEE, December 2018.

29. G. Wadhwa, M. Mathur, M. Todorova, D. A. Omondiagbe, S. Veeramani et al., "Machine Learning Classification Techniques for Breast Cancer Diagnosis," in IOP Conference Series: Materials Science and Engineering (Vol. 495, No. 1, p. 012033). IOP Publishing., 2019.
30. H. Asri, H. Mousannif, H. A. Moatassime and T. Noel, "Using machine learning algorithms for breast cancer risk prediction and diagnosis." Procedia Computer Science, 83, 1064-1069, 2016.
31. X. Zhang, D. He, Y. Zheng, H. Huo, S. Li, et al., "Deep Learning Based Analysis of Breast Cancer Using Advanced Ensemble Classifier and Linear Discriminant Analysis," IEEE Access, vol. 8, pp. 120208–120217, 2020.
32. X. Guo, Z. Liu, C. Sun, L. Zhang, Y. Wang et al., "Deep learning radiomics of ultrasonography: Identifying the risk of axillary non-sentinel lymph node involvement in primary breast cancer." EBioMedicine, 60, 103018, 2020.
33. T. Pang, J. Hsiu, D. Wong, W. Lin and C. Seng, "Expert Systems with Applications Deep learning radiomics in breast cancer with different modalities: Overview and future." Expert Systems with Applications, vol. 158, p. 113501, 2020.
34. L. Li, Q. Feng and X. Wang. "PreMSIm: An R package for predicting microsatellite instability from the expression profiling of a gene panel in cancer." Computational and structural biotechnology journal, 18, 668-675, 2020.
35. X. Zhang, D. He, Y. Zheng, H. Huo, S. Li et al., "Deep learning based analysis of breast cancer using advanced ensemble classifier and linear discriminant analysis." IEEE Access, 8, 120208-120217, 2020.
36. X. Yang, L. Wu, W. Ye, K. Zhao, Y. Wang et al., "Deep Learning Signature Based on Staging CT for Preoperative Prediction of Sentinel Lymph Node," Academic Radiology, no. 4, pp. 1–8, 2019.
37. A. H. Shayma'a, M. S. Sayed, M. I. Abdalla and M. A. Rashwan, "Breast cancer masses classification using deep convolutional neural networks and transfer learning." Multimedia Tools and Applications, 79(41), 30735-30768, 2020.
38. D. Gu, K. Su, and H. Zhao, "A case-based ensemble learning system for explainable breast cancer recurrence prediction," Artificial Intelligence in Medicine, vol. 107, no. April, p. 101858, 2020.
39. Kaggle, "Breast Cancer Prediction Dataset | Kaggle." [Online]. Available: https://www.kaggle.com/merishnasuwal/breast-cancer-prediction-dataset. [Accessed: 11-Mar-2021].
40. L. Tuggener, M. Amirian, K. Rombach, S. Lörwald, A. Varlet, "Automated machine learning in practice: state of the art and recent results." In 2019 6th Swiss Conference on Data Science (pp. 31-36). IEEE, June 2019.
41. Mahesh, B., "Machine Learning Algorithms-A Review." International Journal of Science and Research (IJSR), 9, 381-386, 2020.
42. L. Wang, "Support vector machines: theory and applications" in Springer Science & Business Media, Vol. 177, Ed. 2005.
43. H. Sharma and S. Kumar, "A survey on decision tree algorithms of classification in data mining." International Journal of Science and Research, 5(4), 2094-2097, 2016.
44. Y. Yang, J. Li and Y. Yang, "The research of the fast SVM classifier method." in 2015 12th international computer conference on wavelet active media technology and information processing (ICCWAMTIP) (pp. 121-124). IEEE, December 2015.
45. L. H. Schwamm, R. G. Holloway, P. Amarenco, H. J. Audebert, T. Bakas et al., "A review of the evidence for the use of telemedicine within stroke systems of care: A scientific statement from the American heart association/American stroke association," Stroke, vol. 40, no. 7. pp. 2616–2634, July 2009.
46. A. A. Ibrahim, A. I. Hashad and N. E. M. Shawky, "A comparison of open source data mining tools for breast cancer classification." In Handbook of Research on Machine Learning Innovations and Trends (pp. 636-651). IGI Global, 2017.
47. C. Y. J. Peng, K. L. Lee and G. M. Ingersoll, "An introduction to logistic regression analysis and reporting," The Journal of Educational Research, vol. 96, no. 1, pp. 3–14, 2002.
48. L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, Oct. 2001.
49. T. O. Ayodele, "Types of machine learning algorithms." New advances in machine learning, 3, 19-48, 2010.
50. S. B. Imandoust and M. Bolandraftar, "Application of k-nearest neighbor (knn) approach for predicting economic events: Theoretical background." International Journal of Engineering Research and Applications, 3(5), 605-610, 2013.